Charting Reproducibility and Performance

LLMs in Multilingual Toxic Speech Detection

Bastián González-Bustamante
Leiden University
✉ b.a.gonzalez.bustamante@fgga.leidenuniv.nl

Presentation delivered at the 7th Annual COMPTEXT Conference
University of Vienna, Austria, April 24-26, 2025

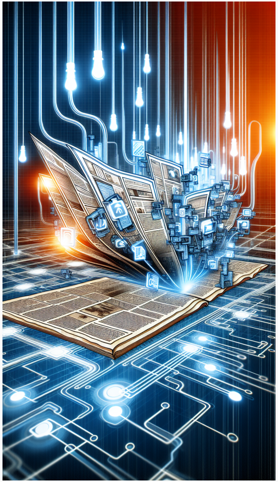
Introduction

The Challenge of Online Toxicity

Digital platforms and incivility. The rise of digital platforms has seen a surge in incivility and toxicity in online interactions, particularly in political discussions.

Need for automated solutions. The sheer volume of online content makes manual annotation impractical, necessitating automated solutions for analysing toxicity and incivility.

LLMs as a potential solution. LLMs offer a promising avenue for automating the annotation process, potentially revolutionising how we study online political discourse.



Artwork by DALL·E 3 model

Introducing LLMs to the Field

Traditional text analysis. Traditional approaches to analysing text data in social sciences often relied on manual coding, dictionaries, or topic modelling, which can be time-consuming and resource-intensive.

The power of LLMs. LLMs, with their zero- and few-shot learning capabilities, offer a powerful alternative, enabling researchers to process and analyse vast amounts of text data efficiently.

Transforming research. This shift towards AI-powered analysis has the potential to transform computational social science research, particularly in understanding online toxicity and incivility.

OpenAI's GPTs: A Closer Look

GPT and research

Researchers have increasingly utilised OpenAI's GPT models through their **API for various tasks**, including text annotation and classification.

Advantages of GPT models

GPT models offer several advantages, including **ease of use**, access to substantial computational resources, and **impressive performance** in many tasks.

Concerns and limitations

However, concerns remain regarding **reproducibility**, data privacy, and the potential for bias in these proprietary models.

The Rise of Open-Source LLMs

Addressing the limitations

Open-source LLMs have emerged as a response to concerns surrounding proprietary models, emphasising transparency, **reproducibility**, and community-driven development.

Performance and potential

Recent studies have shown that open-source LLMs can match or **even surpass** the performance of proprietary models in specific text annotation tasks, particularly in zero-shot classification.

Cost-effective alternatives

Some open-source "Small Language Models" (SLMs) offer comparable performance with significantly **reduced computational requirements** and costs.

Data and Methods

Data and Ground Truth

Arabic

5000 tweets manually annotated

Chinese

5000 messages for toxic detection

English

5000 Wikipedia comments

German

5000 Twitter and Facebook comments

Hindi

5000 Twitter and Facebook comments

Russian

5000 comments on social network OK

Spanish

5000 messages for toxic detection

Note. We split the samples in a proportion of 70/15/15 for training, validation, and testing in case of fine-tuning jobs. The samples correspond to ground-truth data prepared for CLEF TextDetox (2024).

LLMs for Zero-Shot Classification

Our **prompt strategy** was based on the core definitions of Perspective and Google for toxicity with the following system message:

System message

Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC:** Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC:** Civil or nice comments that are unlikely to discourage conversation-

Along with providing texts of our balanced sample, we also listed the categories for the task as follows: **"Respond with only the category (TOXIC or NONTOXIC). Do not provide any additional analysis or explanation"**.

LLMs for Zero-Shot Classification

89 LLMs

Run **776 times** under different conditions (e.g., parameters, API/local, temperature, datasets/language) for

(1) **error rate analysis**
(2) **meta-analysis**

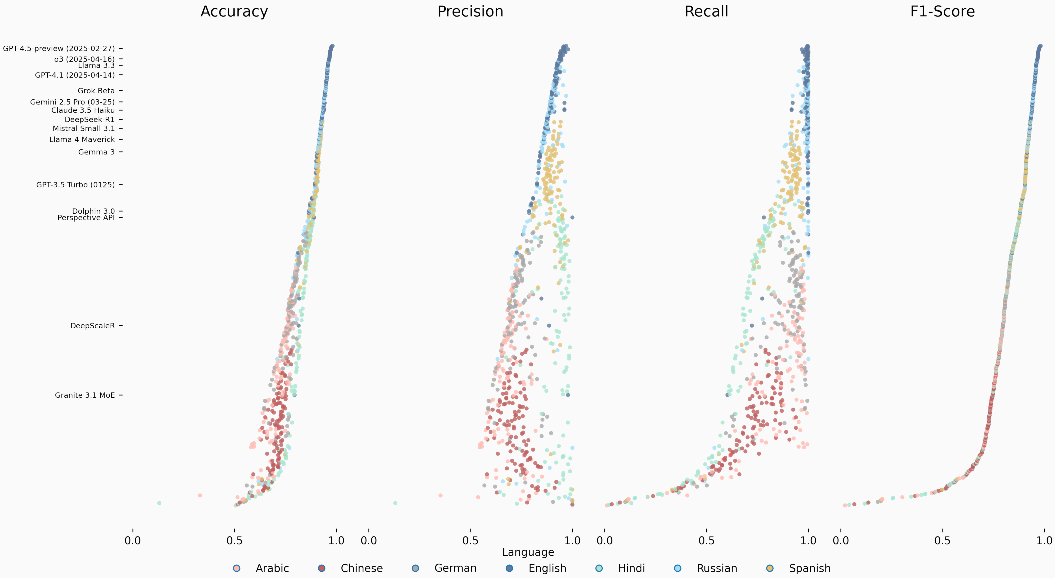
SOTA closed-source LLMs

o3, o4-mini, GPT-4.1, GPT-4.5-preview, Gemini 2.5, Grok 3 Beta, Claude 3.7 Sonnet

SOTA open-source LLMs

Llama 4 Maverick (400B) and Scout (107B), Mistral 3.1 (24B), Llama 3.3 (70B), DeepSeek-R1 (671B), DeepSeek-V3 (671B)

Results



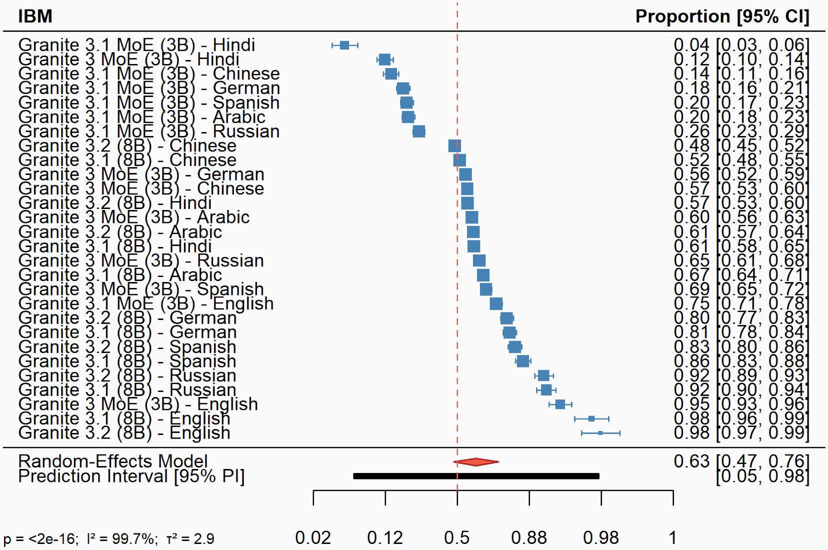
Note. 755 models have been plotted, highlighting the highest F1 for some. Perspective API has been incorporated descriptively, not in the meta-analysis.

Error Rate Analysis Overview (Top-15 LLMs)

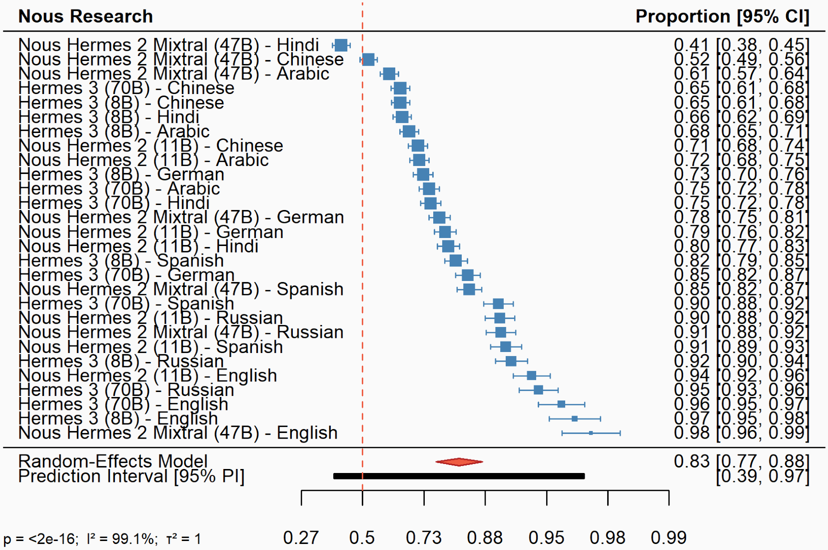
General Pool / 🇬🇧 English	Accuracy	Precision	Recall	F1	Deploy	Provider
Granite 3.2 (8B)	0.981	0.969	0.995	0.982	Local	IBM
Nous Hermes 2 Mixtral (47B)	0.976	0.957	0.997	0.977	Local	Nous Research
Granite 3.1 (8B)	0.976	0.959	0.995	0.976	Local	IBM
OLMo 2 (7B)	0.975	0.954	0.997	0.975	Local	AllenAI
GPT-4.5-preview (2025-02-27)	0.973	0.956	0.992	0.974	API	OpenAI
Yi Large	0.973	0.978	0.968	0.973	API	01 AI
Command R7B Arabic (7B)	0.972	0.959	0.987	0.972	Local	Cohere
Yi 1.5 (34B)	0.971	0.951	0.992	0.971	Local	01 AI
Mistral OpenOrca (7B)	0.969	0.942	1.000	0.970	Local	Mistral
Hermes 3 (8B)	0.969	0.961	0.979	0.970	Local	Nous Research
Phi-3 Medium (14B)	0.969	0.966	0.973	0.969	Local	Microsoft
GPT-4 (0613)	0.968	0.940	1.000	0.969	API	OpenAI
GLM-4 (9B)	0.968	0.942	0.997	0.969	Local	Zhipu AI
DeepSeek-V3 (671B)	0.968	0.944	0.995	0.969	API	DeepSeek-AI
Sailor2 (20B)	0.968	0.944	0.995	0.969	Local	Sailor2

Private-closed; **Open-source**

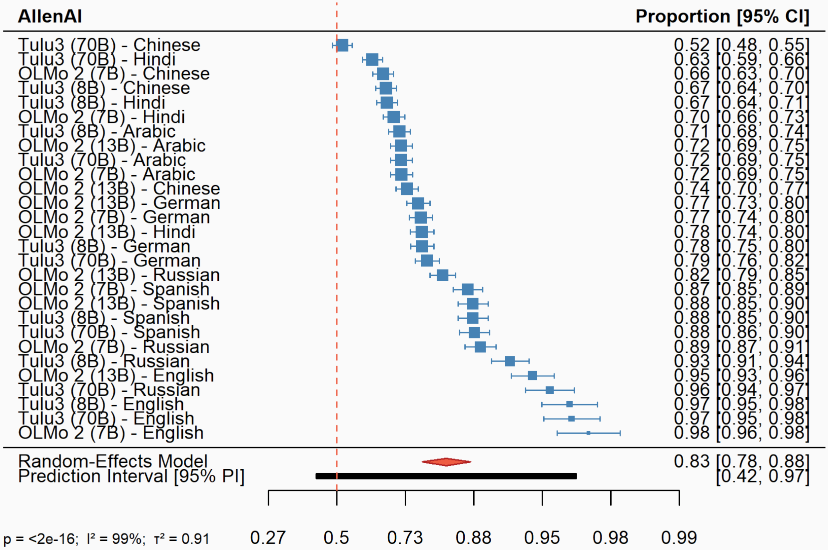
Forest Plot Random-Effects IBM LLMs



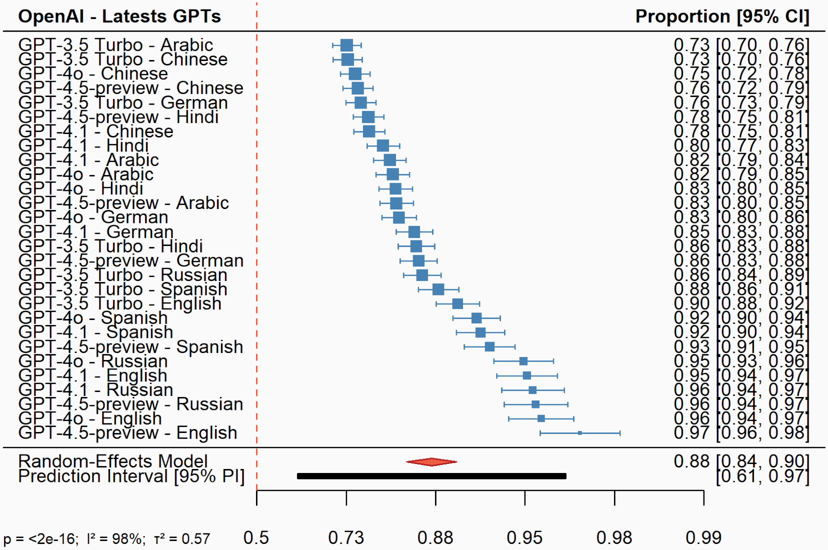
Forest Plot Random-Effects Nous Research LLMs



Forest Plot Random-Effects AllenAI LLMs



Forest Plot Random-Effects OpenAI GPTs



Error Rate Analysis per Language (Top-15 LLMs)

Arabic	Accuracy	F1
GPT-4.5-preview (2025-02-27)	0.800	0.827
GPT-4o (2024-11-20)	0.787	0.821
GPT-4.1 (2025-04-14)	0.780	0.816
GPT-4o (2024-05-13)	0.779	0.816
GPT-4 Turbo (2024-04-09)	0.780	0.815
GPT-4o (2024-08-06)	0.768	0.809
GPT-4 (0613)	0.784	0.808
Gemini 1.5 Pro	0.763	0.804
Gemini 2.5 Pro (03-25)	0.760	0.803
GPT-4.1 mini (2025-04-14)	0.771	0.802
GPT-4.1 nano (2025-04-14)	0.796	0.801
Gemini 1.5 Flash-8B	0.784	0.801
Grok 3 Mini Beta	0.763	0.801
Grok 3 Mini Fast Beta	0.763	0.801
Aya Expanse (32B)	0.765	0.800

Chinese	Accuracy	F1
GPT-4.1 (2025-04-14)	0.776	0.779
GPT-4o (2024-05-13)	0.771	0.778
Grok 3 Fast Beta	0.756	0.775
GPT-4o (2024-08-06)	0.764	0.773
Grok 3 Beta	0.749	0.771
Grok 2 (1212)	0.729	0.762
GPT-4 Turbo (2024-04-09)	0.747	0.761
Grok Beta	0.748	0.760
Gemini 1.5 Pro	0.729	0.756
GPT-4.5-preview (2025-02-27)	0.752	0.756
Gemini 2.0 Flash	0.728	0.752
Gemini 1.5 Flash	0.715	0.752
Gemini 2.5 Pro (03-25)	0.725	0.751
GPT-4o (2024-11-20)	0.755	0.751
DeepSeek-R1 (671B)	0.717	0.747

Private-closed; **Open-source**

Error Rate Analysis per Language (Top-15 LLMs)

German	Accuracy	F1
GPT-4.5-preview (2025-02-27)	0.857	0.859
GPT-4.1 (2025-04-14)	0.843	0.854
Hermes 3 (70B)	0.845	0.848
GLM-4 (9B)	0.829	0.844
Qwen 2.5 (32B)	0.829	0.843
GPT-4 (0613)	0.829	0.841
Grok 3 Mini Beta	0.816	0.836
Grok 3 Mini Fast Beta	0.813	0.835
GPT-4o (2024-08-06)	0.815	0.835
OpenThinker (32B)	0.816	0.834
GPT-4o (2024-05-13)	0.815	0.833
DeepSeek-R1 D. Qwen (14B)	0.823	0.831
GPT-4o (2024-11-20)	0.813	0.831
Gemini 1.5 Flash-8B	0.812	0.831
Aya (32B)	0.813	0.830

Hindi	Accuracy	F1
Mistral Saba (24B)	0.900	0.895
Gemma 2 (9B)	0.889	0.890
Gemma 3 (4B)	0.879	0.885
Grok 2 (1212)	0.888	0.884
Llama 4 Maverick (400B)	0.891	0.883
Gemma 3 (27B)	0.887	0.880
Grok 3 Beta	0.884	0.877
Llama 3.1 (405B)	0.883	0.876
Grok 3 Fast Beta	0.883	0.875
Gemini 1.5 Flash	0.884	0.874
Mistral Small (22B)	0.865	0.871
Pixtral Large (2411)	0.876	0.870
Gemini 1.5 Pro	0.876	0.866
Gemini 2.0 Flash-Lite (001)	0.877	0.865
Gemini 2.0 Flash-Lite (02-05)	0.877	0.865

Private-closed; **Open-source**

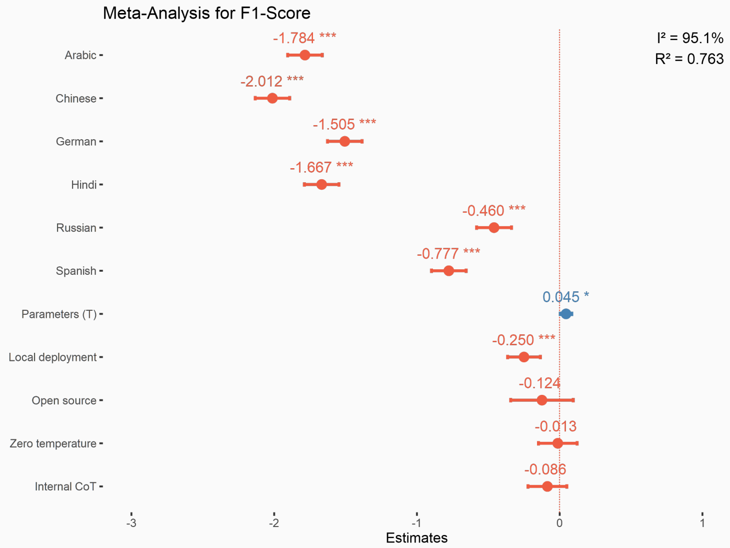
Error Rate Analysis per Language (Top-15 LLMs)

Russian	Accuracy	F1
GPT-4.1 nano (2025-04-14)	0.961	0.962
Claude 3.5 Sonnet	0.957	0.958
Claude 3.7 Sonnet	0.957	0.958
GPT-4.5-preview (2025-02-27)	0.956	0.958
Tulu3 (70B)	0.957	0.957
GPT-4.1 (2025-04-14)	0.955	0.956
QwQ (32B)	0.953	0.954
GPT-4o (2024-11-20)	0.949	0.952
GPT-4o (2024-05-13)	0.948	0.951
Command R7B Arabic (7B)	0.951	0.950
Gemini 1.5 Flash-8B	0.948	0.950
GLM-4 (9B)	0.948	0.950
GPT-4 (0613)	0.947	0.949
Qwen 2.5 (32B)	0.947	0.949
DeepSeek-V3 (671B)	0.947	0.949

Spanish	Accuracy	F1
GPT-4.5-preview (2025-02-27)	0.932	0.931
Athene-V2 (72B)	0.925	0.925
GPT-4.1 (2025-04-14)	0.923	0.924
Qwen 2.5 (72B)	0.924	0.923
GPT-4o (2024-05-13)	0.921	0.923
GPT-4o (2024-11-20)	0.921	0.921
GPT-4 (0613)	0.920	0.919
Grok Beta	0.916	0.917
Pixtral Large (2411)	0.913	0.917
OpenThinker (32B)	0.916	0.916
Qwen 2.5 (14B)	0.915	0.916
GPT-4 Turbo (2024-04-09)	0.912	0.916
GPT-4o (2024-08-06)	0.913	0.915
Qwen 2.5 (32B)	0.915	0.914
Gemini 2.0 Flash	0.909	0.914

Private-closed; Open-source

Random-Effects Pooled Model



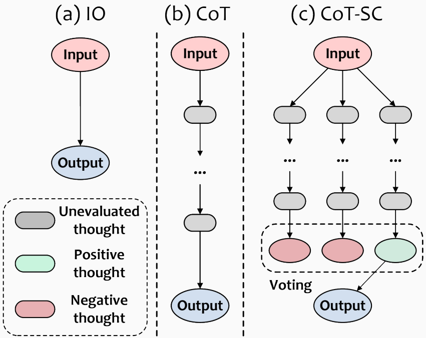
Ancillary Findings

Chain-of-Thought and Reasoning Models

Although we are not testing CoT, we have tested different **reasoning models**:

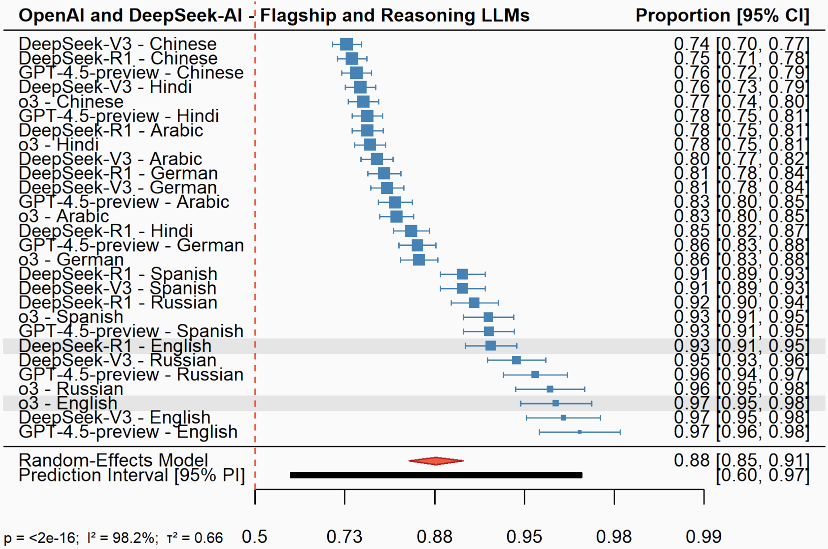
- o1, o1-mini, o3, o3-mini, o4-mini, DeepSeek R-series (including distilled LLMs based on Llama or Qwen), Gemini thinking models, Marco-o1-CoT.

These models run an *internal* CoT that emulates the prompting technique.

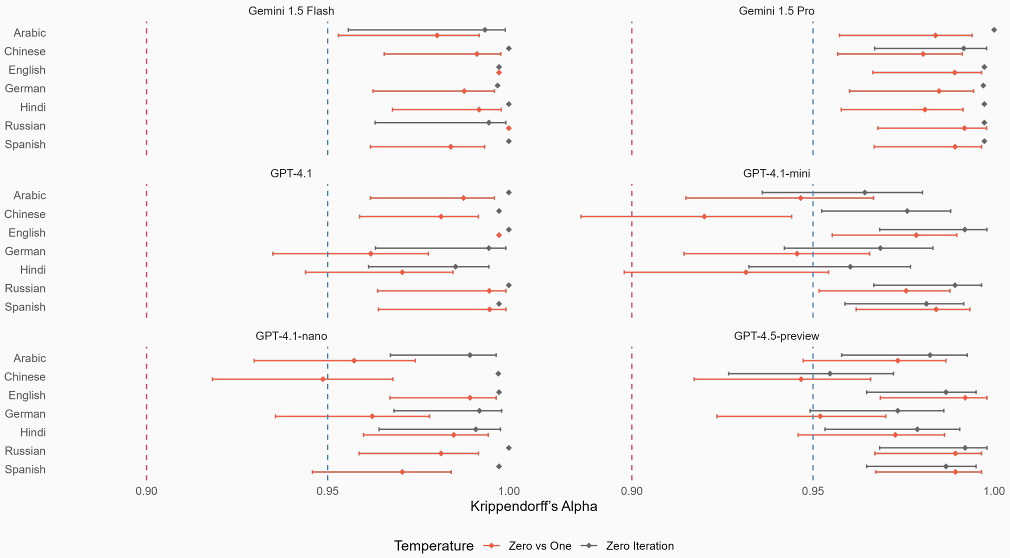


Source: Din et al. (2024).

Forest Plot Random-Effects Reasoning Models



Temperature Experiments



Takeaways

Takeaways

- Although SOTA LLMs outperform other LLMs in zero-shot classification and traditional approaches, there is a relevant **performance heterogeneity between languages**.
- Classifying texts in English shows an outstanding performance. Although open-source models seem to predominate in English classification, robustness checks with **two-way interactions between languages and open-source models did not show significance**.
- Reasoning models with **internal CoT do not perform better than other LLMs**. Maybe "overthinking" in this task does not pay off.
- Research has indicated that only open-source LLMs at zero temp. offer full reproducibility. However, some **SOTA private LLMs** not only show good performance but also **high levels of reliability** (not full inter-coder agreement).

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some ☕
Download the latest version from GitHub